United States Patent [19]

Schmitt

[11] **Patent Number: 5,062,143**

[45] **Date of Patent: Oct. 29, 1991**

[54] **TRIGRAM-BASED METHOD OF LANGUAGE IDENTIFICATION**

[75] Inventor: **John C. Schmitt,** Indialantic, Fla.

[73] Assignee: **Harris Corporation,** Melbourne, Fla.

[21] Appl. No.: **485,115**

[22] Filed: **Feb. 23, 1990**

[51] **Int. Cl.**[5] **G06K 9/62;** G06K 9/72

[52] **U.S. Cl.** **382/36;** 382/39; 382/40

[58] **Field of Search** 382/36, 40, 9, 37, 38, 382/39

[56] **References Cited**

U.S. PATENT DOCUMENTS

3,969,698	7/1976	Bollinger et al.	381/43
4,754,489	6/1988	Bokser et al.	382/40
4,829,580	5/1989	Church	381/52

Primary Examiner—Leo H. Boudreau
Assistant Examiner—Steven P. Fallon
Attorney, Agent, or Firm—Evenson, Wands, Edwards, Lenahan & McKeown

[57] **ABSTRACT**

A mechanism for examining a body of text and identifying its language compares successive trigrams into which the body of text is parsed with a library of sets of trigrams. For a respective language-specific key set of trigrams, if the ratio of the number of trigrams in the text, for which a match in the key set has been found, to the total number of trigrams in the text is at least equal to a prescribed value, then the text is identified as being possibly written in the language associated with that respective key set. Each respective trigram key set is associated with a respectively different language and contains those trigrams that have been predetermined to occur at a frequency that is at least equal to a prescribed frequency of occurrence of trigrams for that respective language. Successive key sets for other languages are processed as above, and the language for which the percentage of matches is greatest, and for which the percentage exceeded the prescribed value as above, is selected as the language in which the body of text is written.

6 Claims, 2 Drawing Sheets